ZEROFOX INTELLIGENCE

Flash

AI Ransomware Toolkit Automates Operations

F-2026-06-16a

Classification: TLP:CLEAR

Criticality: Low

Intelligence Requirements: AI, Ransomware, Threat Actor

June 16, 2026

Scope Note

ZeroFox Intelligence is derived from a variety of sources, including—but not limited to—curated open-source accesses, vetted social media, proprietary data sources, and direct access to threat actors and groups through covert communication channels. Information relied upon to complete any report cannot always be independently verified. As such, ZeroFox applies rigorous analytic standards and tradecraft in accordance with best practices and includes caveat language and source citations to clearly identify the veracity of our Intelligence reporting and substantiate our assessments and recommendations. All sources used in this particular Intelligence product were identified prior to 4:00 AM (EDT) on June 15, 2026; per cyber hygiene best practices, caution is advised when clicking on any third-party links.

| Flash | AI Ransomware Toolkit Automates Operations

| Key Findings

- On June 2, 2026, security researchers discovered an unknown threat actor was almost certainly using commercially available artificial intelligence (AI) technologies to develop and iteratively test Endpoint Detection and Response (EDR) evasion techniques within a post-exploitation framework that was presented as a “red team” exercise.
- The threat actor reportedly used AI to accelerate tool development and testing, but the operation remained human-driven. AI was very likely used primarily to coordinate workflows and support experimentation, while the EDR-bypass work followed a structured engineering test cycle that included human review and iteration.
- ZeroFox assesses the framework was very likely built for criminal use rather than legitimate security testing. The activity is linked to known ransomware deployment and data theft operations, and the red team framing was likely a pretext to circumvent the AI model's safety guardrails.
- ZeroFox assesses that the use of AI to accelerate tooling and test evasion techniques likely lowers the barrier to entry for sophisticated, red team-style

intrusions but does not change defensive priorities. Fundamentals such as timely patching, multi-factor authentication (MFA), modern authentication (such as passkeys), and broad EDR deployment likely remain the primary mitigations.

Details

On June 2, 2026, security researchers discovered an unknown threat actor almost certainly using commercially available AI technologies to develop and iteratively test EDR evasion techniques within a post-exploitation framework presented as a red team exercise.[1]

The activity surfaced when an anomalous endpoint registered within a monitored customer tenant triggered alerts for malicious payloads originating from a user's directory on the target system. The files in the target directory indicated a broader detection evasion framework comprising four core components:

- Cobalt Strike profiles engineered to make beacon traffic resemble legitimate web requests;
- Telegram bot Application Programming Interface (API)–based command-and-control (C2) mechanisms that routed communications through Telegram infrastructure;
- Python-based scripts for injecting shellcode into legitimate Windows executables while preserving their original functionality; and
- Cloudflare Worker acting as a front-end redirector to conceal the actual backend C2 server.[2]

According to the security researchers' report, a Git repository held a framework aligned with two functions: an automated Active Directory (AD) discovery panel and a laboratory that iteratively developed and tested malware against multiple EDR agents. The AD discovery panel most closely resembles automated, AI-driven functionality but reportedly does not represent an autonomously reasoning large language model (LLM);

[1] hXXps://www.sophos[.]com/en-us/blog/pointing-a-cursor-at-evading-detection
[2] hXXps://www.bleepingcomputer[.]com/news/security/ai-built-ransomware-toolkit-automates-edr-evasion-ad-discovery

rather, it collects observations from completed tasks, selects the next action from a predefined set, dispatches work to remote agents, and re-evaluates as results return.

The security researchers' report indicated the threat actor used a Virtual Machine (VM) and the AI-native Integrated Development Environment (IDE) cursor to build EDR-bypass tooling capable of creation, testing, analysis, and refinement. Multiple AI agents were reportedly configured with defined roles: one agent running the Claude Opus 4.5 model coordinated core operations and set rules for the others, an additional agent tested tooling against EDR products, and the remaining agents handled operational security hardening, documentation, proxy stress-testing, and VM deployment. Further, agent-generated code issues and commits were reportedly relayed to Git via the Model Context Protocol (MCP).

The test environment reportedly comprised separate Windows Server 2022 VMs—one per EDR product under test and a control with no EDR installed—and a fourth VM running a post-exploitation C2 server on Ubuntu. The tooling was tested in VMs against multiple commercial EDR products, including Microsoft Windows Defender.

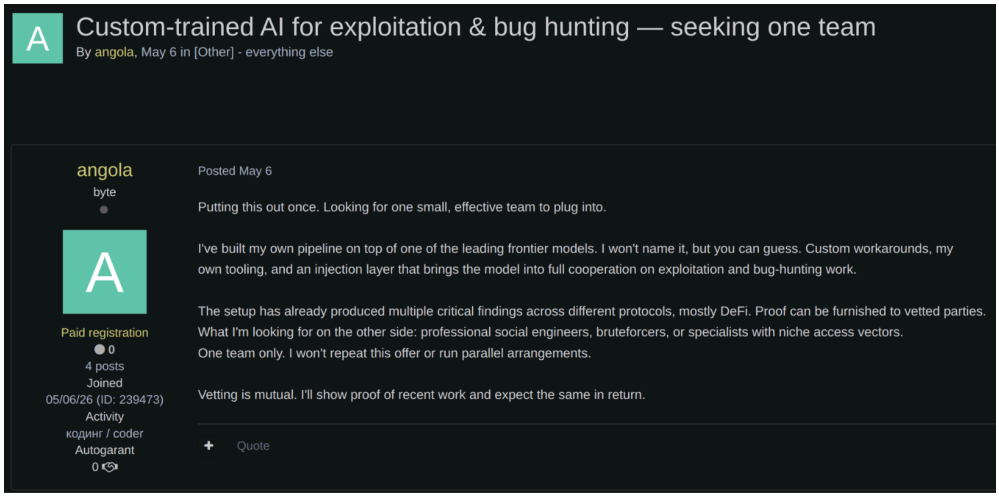
- A Python-based payload generator at the core of the framework produced roughly 80 modules testing more than 70 techniques, wrapping payloads (mostly written in Rust and Go) in layers of encryption and providing alternative execution methods intended to resist sandboxing and EDR detection.

While AI agents initially reported a high failure rate, the modules were considered almost universally effective against the EDR agents after iteration. However, the documented test output does not necessarily support that success claim, and the reason for the discrepancy is unclear.[3] Multiple Python scripts on the host were partially AI-generated and written in Russian, but there is no other evidence available that indicates nationality or attribution.[4]

Threat actors are almost certainly integrating commercial AI coding assistants and agent orchestration into offensive tool development. ZeroFox has observed threat actors advertising similar custom-built, AI-enabled malware tooling on dark web forums.

[3] hXXps://www.darkreading[.]com/endpoint-security/attackers-automate-edr-evasion-testing
[4] hXXps://www.bleepingcomputer[.]com/news/security/ai-built-ransomware-toolkit-automates-edr-evasion-ad-discovery

- In May 2026, unvetted threat actor “angola” advertised a custom AI for exploitation and bug hunting. Due to customer complaints on the forum regarding AI malware, it is unlikely the tool will be advertised again; however, the post demonstrates that threat actors are almost certainly experimenting with ways to monetize AI-made tools.



angola’s post on Exploit
Source: ZeroFox Intelligence

Despite AI involvement at multiple stages in the unidentified threat actor’s development of EDR evasion techniques observed by security researchers, the workflow very likely remained human-driven. AI likely compressed development and testing timelines rather than autonomously conducting the operation; the available evidence does not support claims of autonomous AI capability. The documented gap between the AI agents' self-reported bypass success and the actual test output very likely underscores the need for caution when interpreting AI-reported outcomes.

- In this instance, the framework was very likely developed for criminal rather than legitimate purposes. The observed activity's links to known ransomware and data theft operations, together with Cobalt Strike operator logs that referenced a ransom note and organizations listed on a ransomware data leak site, almost certainly indicate criminal intent.

- The use of a red team pretext to circumvent AI safety guardrails is consistent with previously documented patterns of AI misuse, in which defensive or research

framing is used to elicit otherwise-restricted assistance.[5] This is a recurring tradecraft theme rather than a novel one, and it is likely to persist as agentic AI tooling matures.

It is very likely that other threat actors will adopt similar AI-assisted development and testing workflows, given the demonstrated reduction in effort required to produce and refine evasion tooling. ZeroFox recommends that organizations sustain defense-in-depth techniques: timely patching, MFA, modern authentication mechanisms such as passkeys, broad and well-tuned EDR deployment, and monitoring for the behaviors associated with Cobalt Strike and AD-discovery activity.

[5] hXXps://www.infosecurity-magazine[.]com/news/ai-edr-evasion-tooling/

Recommendations

- Develop a comprehensive incident response strategy.
- Deploy a holistic patch management process, and ensure all IT assets are patched with the latest software updates as quickly as possible.
- Adopt a Zero-Trust cybersecurity architecture based upon a principle of least privilege.
- Implement network segmentation to separate resources by sensitivity and/or function.
- Ensure critical, proprietary, or sensitive data is always backed up to secure, off-site, or cloud servers at least once per year—and ideally more frequently.
- Implement secure password policies, phishing-resistant MFA, and unique credentials.
- Configure email servers to block emails with malicious indicators, and deploy authentication protocols to prevent spoofed emails.
- Proactively monitor for compromised accounts and credentials being brokered in deep and dark web (DDW) forums.
- Leverage cyber threat intelligence to inform the detection of relevant cyber threats and associated tactics, techniques, and procedures (TTPs).

Appendix A: Traffic Light Protocol for Information Dissemination

	Red	Amber
WHEN SHOULD IT BE USED?	**Sources may use** TLP:RED when information cannot be effectively acted upon by additional parties and could lead to impacts on a party's privacy, reputation, or operations if misused.	**Sources may use** TLP:AMBER when information requires support to be effectively acted upon but carries risks to privacy, reputation, or operations if shared outside of the organizations involved.
HOW MAY IT BE SHARED?	**Recipients may NOT share** TLP:RED with any parties outside of the specific exchange, meeting, or conversation in which it is originally disclosed.	**Recipients may ONLY share** TLP:AMBER information with members of their own organization and its clients, but only on a need-to-know basis to protect their organization and its clients and prevent further harm. **Note that** TLP:AMBER+STRICT restricts sharing to the organization only.

	Green	Clear
WHEN SHOULD IT BE USED?	**Sources may use** TLP:GREEN when information is useful for the awareness of all participating organizations, as well as with peers within the broader community or sector.	**Sources may use** TLP:CLEAR when information carries minimal or no risk of misuse in accordance with applicable rules and procedures for public release.
HOW MAY IT BE SHARED?	**Recipients may share** TLP:GREEN information with peers and partner organizations within their sector or community but not via publicly accessible channels.	**Recipients may share** TLP:CLEAR information without restriction, subject to copyright controls.

Appendix B: ZeroFox Intelligence Probability Scale

All ZeroFox intelligence products leverage probabilistic assessment language in analytic judgments. Qualitative statements used in these judgments refer to associated probability ranges, which state the likelihood of occurrence of an event or development. Ranges are used to avoid a false impression of accuracy. This scale is a standard that aligns with how readers should interpret such terms.

Almost No Chance	Very Unlikely	Unlikely	Roughly Even Chance	Likely	Very Likely	Almost Certain
1-5%	5-20%	20-45%	45-55%	55-80%	80-95%	95-99%